
DISPLAYING KHMER

The purpose of this document is to help developers who do not know the Khmer script to understand what is involved in displaying Khmer Unicode correctly.

The Script

Without getting too much into the intricacies of the language, and asking in advance for excuses from Khmer linguists for generalizations and lack of scientific methodology, here are some of the particularities of Khmer Script writing:

- It is written from left to right, with characters being placed also above and below the main line of writing.
- In Khmer words are not separated by spaces. A space in Khmer is a punctuation sign similar to a comma (slightly stronger than a comma for some linguists).

ជំរិតគឺជាផ្នែកមួយនៃព្រាងនិទានរបស់ខ្មែរយើង

- A word is composed of clusters, sometimes also called syllemes. They are not a proper syllable, as syllables are a unit of consonants and vowels pronounced in one stroke of breath. Consonants pronounced after a vowel are part of the syllable, but not part of the cluster or sylleme.

This is a sylleme: ធ្វើ

- In Khmer, consonants have two forms: 1) A “normal” one, written in the main line of text, and 2) A subscript one, placed under another consonant. A consonant subscript placed under a normal consonant is read after the normal consonant, without any vowel sound between them (in theory).

Some normal consonants with subscripts: ម្ម ឱ ណ្ណ

Some of the subscript consonants are placed to only occupy space under the normal consonant, but others extend (because of their shape) to be placed under and before the main consonant, or under and after.

Examples: ស្រ ផ្ស

- There are other signs called independent vowels that behave pretty much like consonants (even if many of them have vowels sounds) and for typographical purposes they are considered almost as consonants (except that they may not have vowels in their syllemes).
- Each vowel in Khmer has a specific location where it has to be placed in reference to the main consonant of the sylleme. If the vowel has only one graph, it may be placed below the consonant (in which case it also goes below any subscript consonants, excepting subscript RO,

in which case the vowel it may be placed right of lower part of the subscript RO, see third example), left of (before) the consonant, above the consonant, or right of it (behind). There are vowels that have two graphs that need to be placed. Some of these consonants have to be placed before and above the consonant, below and after, below and above or before and after the consonant.

Here are some examples, using as a base the consonant ក្រ :

ក្រ, ក្រ្ម, ក្រ្យ, ក្រ្រ, ក្រ្ល, ក្រ្វ, ក្រ្ម្រ, ក្រ្យ្រ, ក្រ្រ្រ

- There are also in Khmer some diacritical signs that for part of syllemes. In some special cases, one of these diacritical signs may change shape and location, depending on which vowels follows it.

Examples of diacriticals:

ន្រ្ទ, ប្រ្ទ, រ្រ្ទ, ន្រ្ទ also ន្រ្ទ្រ, but when followed by vowel sign ្រ turns to ន្រ្ទ្រ

- There is one ligature in Khmer that must be used. When the consonant ឃ is followed by the vowel ្រ, the consonant changes its shape, giving: ឃ្រ

A Khmer sylleme is then always composed of:

- A normal consonant or Independent Vowel (one and only one).
- Zero, one or two subscript Consonants or Independent Vowels
- Zero, one or two diacritic signs
- Zero or one vowel.

Unicode

Some specificities of Unicode Khmer codepoints:

- Unicode defines one codepoint (one number) for each normal consonant, but not for subscript consonants, subscripts consonants are constructed by writing first the Unicode codepoint u+17D2 and then the codepoint of the normal consonant.
- Some vowels that include the sign ្រ in their glyph (in their shape) do not have Unicode codepoints. The following vowels (as they are studied by Khmer students and frustrated foreigners) have to be constructed using two Unicode codepoints (consonant ក្រ is used as a base):

ក្រ្រ្រ, ក្រ្រ្រ្រ, ក្រ្រ្រ្រ្រ

- Also, the vowel ្រ្រ has to be constructed using the codepoints for vowels ្រ and ្រ្រ

Unicode Khmer is always typed in the following order:

- Consonant or Independent Vowel.
- Subscript Consonants or subscript Independent Vowels (if any)
- Diacritic signs (if any)
- Vowel (if any).

This order is different from the order in which Khmer is traditionally written, as normally preceding vowels are written before, but it is much better for sorting (because the main consonant, written before, commands sorting) and for speech rendering.

Rendering

In order to make sure that you can render Khmer correctly, you must assure that:

- Normal subscript consonants are placed correctly, including the case for two subscripts (and the specific case of the second subscript being subscript RO). Assure that subscript consonants that occupy space at the right or the left of the main consonant actually have this space (when subscript RO [u+17D2 u+179A] is typed, the main consonant has to move to the right to allocate space for the subscript RO). If there are two subscript consonants, and one of them is subscript RO, subscript RO has to change shape in order to accommodate the other subscript consonant.

Correct: ម្មស្រជ្ជុណ្ណ

Characters: ម ្ម ស ្រ ជ ្ជ ុ ណ ្ណ

Unicode [u+1798 u+17d2 u+179f u+17d2 u+179a u+1795 u+17d2 u+179f u+1781 u+17d2 u+1798 u+178e u+17d2 u+178a]

Correct: ក្រៃ

Characters: ក ្រ ៃ

Unicode [u+1780 u+1793 u+17d2 u+17178f u+17d2 u+179a u+17c3]

Correct: អង្គស

Characters: អ ង ្គ ស

Unicode [u+17a2 u+1784 u+17d2 u+1782 u+17d2 u+179b u+17 c1]

- Vowels are placed correctly, including the ones that are clearly placed before the main consonant (reordering), or a part of them is placed there.

Correct: ស្រលទៅអង្គយដីស្រើឈ្មោះ

Characters: ស ្រ ល ទ ៅ អ ង ្គ យ ដី ស្រើ ឈ្មោះ

Unicode [u+179f u+17d2 u+179a u+17bc u+179b u+1791 u+17c5 u+17a2 u+1784 u+17d2 u+1782 u+17bb

- Ligatures work correctly.

Correct: ប៉ា

Characters: ប ៉ា

Unicode [u+1794 u+17b6]

- Diacritic marks are placed correctly, including the one that changes shape depending on context.

Correct: ន័យទុគ៌ិមប៉ាត់ស្ទើស៊ី

Characters: ន ៉ យ ទ ុ គ ិ ម ប ៉ ា ត ៉ ំ ស ឺ រ ោ ស ឺ ឺ

Unicode [u+1793 u+17d0 u+1799 u+1791 u+17bb u+1782 u+17cc u+1798
u+1794 u+17c9 u+17b6 u+178f u+17cb u+179f u+17ca u+17c4 u+179f u+17ca u+17b8]